第27卷 第4期　工 程 数 学 学 报　Vol. 27 No. 4
2010年08月　CHINESE JOURNAL OF ENGINEERING MATHEMATICS　Aug. 2010

文章编号:1005-3085(2010)04-0571-13

临床试验不依从现象下相关数据的统计推断*

戴　君，张忠占

(北京工业大学应用数理学院，北京　100124)

摘　要: 假定在临床试验中必需考虑受试者服药状态相关性以及受试者协变量对治疗效果产生的影响。本文证明了治疗效果参数在这一假设下依然具有可识别性。利用依从者治疗效果参数的近似极大似然估计，本文构造了此参数的记分检验统计量。模拟试验表明，即使对较小样本，该检验方法犯第一类错误的概率也接近给定的检验水平。

关键词: 不依从现象；工具变量；相关数据；记分检验

分类号: **AMS**(2000) 62P10　　**中图分类号**: O213　　**文献标识码**: A

1　引言

临床试验中出现部分受试者不能完全按照治疗方案接受治疗的现象称为不依从现象。该现象的出现，破坏了随机化试验的模型基础，导致响应变量不能完全反映治疗效果，给治疗效果估计带来困难。考虑受试者协变量对治疗效果有影响这一因素，也使该统计推断问题更加复杂。

定义二分变量 D 表示受试者服药状态，二分变量 Y 表示受试者治疗响应结果。治疗效果表现为同一名受试者服用药物 ($D=1$) 下潜在响应变量 $Y_{(D=1)}$ 与服用安慰剂 ($D=0$) 下潜在响应变量 $Y_{(D=0)}$ 的差异。对于同一名受试者，不能同时观测到潜在治疗响应变量 $Y_{(1)}$ 和 $Y_{(0)}$，只能观测到 $Y=Y_{(1)}\cdot D+Y_{(0)}\cdot(1-D)$，因此不能直接计算每个受试者的治疗效果 $Y_{(1)}-Y_{(0)}$。

Imbens 和 Angrist[1] 提出用工具变量法推断治疗参数。引入工具变量 Z 表示设计试验分组状态，在受试者服用药物或安慰剂前首先对受试者进行分组 (治疗组 $Z=1$，对照组 $Z=0$)。定义二分变量 $D_{(z)}$ 表示在设计试验分组 $Z=z$ 下潜在治疗状态指示变量[2,3]。利用潜在试验状态变量 $D_{(1)}$ 和 $D_{(0)}$ 把总体分为四组。依从者 ($D_{(1)}>D_{(0)}$)，总是服药者 ($D_{(1)}=D_{(0)}=1$)，从不服药者 ($D_{(1)}=D_{(0)}=0$)，违背者 ($D_{(1)}<D_{(0)}$)。试验状态指示变量可以表示为 $D=Z\cdot D_{(1)}+(1-Z)\cdot D_{(0)}$。试验中可以观测到 Z 和 D，但不能同时观测到同一受试者的两个潜在试验状态变量 $D_{(1)}$ 和 $D_{(0)}$，因此我们不能识别个体属于哪一组。定义变量 $Y_{(zd)}$ 表示受试者分组在 z 组、服药状态在 d 组下，受试者潜在治疗响应结果[4,5]。若某个体 $D_{(0)}=0$，将不能观测到此个体的变量 $Y_{(01)}$。

Imbens 和 Angrist[1] 在不考虑协变量对治疗效果有影响的条件下，给出一个简单工具估计量，可以识别依从者的平均治疗效果。Imbens 和 Rubin[6]，Abadie[7,8] 提出在相同假设下可以识别依从者潜在治疗响应的边缘分布。Abadie[9] 进一步研究针对协变量对治疗效果有影响的情况，证明了在给定协变量的条件下，依从者平均治疗效果是可识别的。

收稿日期: 2008-07-14.　**作者简介**: 戴君 (1983年1月生)，女，硕士. 研究方向：应用统计.

***基金项目**: 北京市自然科学基金 (1072003)；国家自然科学基金 (10971007).

以上讨论都限定了受试者在临床试验中相互独立不受彼此干扰这一条件，即假设临床试验中可观测到受试者的数据 $O=(X,Z,D,Y)$ 相互独立，其中 X 表示受试者在随机试验前的协变量，如性别、年龄等基线水平。

而在实际临床试验中还会出现在同一区域(病房、试验中心)内的受试者在服药期间相互影响的现象。比如，同一区域内受试者的依从性可能受到趋同心理的影响。这种相关性对治疗效果的评估存在一定影响，本文将针对此类现象，考虑反映受试者服药变量间存在相关性情况下，治疗效果参数的统计推断问题。

第二节，我们利用工具变量法，将模型扩展到受试者服药状态相关的情况，并证明治疗效果的可识别性。第三节，考虑受试者分为 m 组，第 i 组 n_i 人，$i=1,\cdots,m$，组间受试者服药状态独立，组内受试者服药状态相关的情况。根据治疗效果的可识别性，利用治疗效果参数的极大似然估计，构造记分检验统计量。第四节，在特殊模型Logit模型和受试者服药状态相关性的假定下计算检验统计量。第五节，进行模拟研究，考察检验方法的优劣。

2 相关数据下的识别问题

以下假设是针对解决识别问题而提出，应用于很多文章中。

假设2.1 (i) 工具变量的独立性：在给定 X 的条件下，随机向量 $(Y_{(00)},Y_{(01)},Y_{(10)},Y_{(11)},D_{(0)},D_{(1)})$ 与 Z 独立；

(ii) 工具变量的不相容性：$P(Y_{(1d)}=Y_{(0d)}\,|\,X)=1$，对任意的 $d\in\{0,1\}$；

(iii) 第一阶段：$0<P(Z=1\,|\,X)<1$，并且 $P(D_{(1)}=1\,|\,X)>P(D_{(0)}=1\,|\,X)$；

(iv) 单调性：$P(D_{(1)}>D_{(0)}|X)=1$。

若假设2.1成立，则称工具变量 Z 是有效的。

假设2.1的(i)保证试验分组完全随机化。假设2.1的(ii)表示工具变量 Z 只是通过 D 对潜在治疗响应变量作用，即变量对响应的唯一影响是通过在临床试验中的服药状态体现，因此对相同的 X 有 $Y_{(0)}=Y_{(00)}=Y_{(10)}$ 和 $Y_{(1)}=Y_{(01)}=Y_{(11)}$ 成立。假设2.1的(iii)保证在条件 X 下 Z 和 D 相关，并进一步暗示在 $Z=1$ 下 X 的支撑与在 $Z=0$ 条件下 X 的支撑相同。假设2.1的(iv)排除了违背者的存在，并定义总体分别为总是服药者、依从者和从不服药者三组。

接下来，针对 n 个受试者间服药变量 D_i, $i=1,\cdots,n$，相关的情况，分析治疗效果参数的可识别性。

记 $\mathbf{Y}=(Y_1,\cdots,Y_n)$, $\mathbf{D}=(D_1,\cdots,D_n)$, $\mathbf{X}=(X_1,\cdots,X_n)$。

定理2.1 令 $g(\cdot)$ 为 $(\mathbf{Y},\mathbf{D},\mathbf{X})$ 的任意可测实值函数，满足 $E|g(\mathbf{Y},\mathbf{D},\mathbf{X})|<\infty$。定义

$$\kappa_n=1-\sum_{\substack{z_i=0,1\\ i=1,\cdots,n}}\frac{\prod_{i=1}^n Z_i^{z_i}(1-Z_i)^{1-z_i}(1-\prod_{i=1}^n D_i^{z_i}(1-D_i)^{1-z_i})}{P(Z_1=z_1,\cdots,Z_n=z_n\,|\,X_1,\cdots,X_n)},$$

在假设2.1条件下

$$E\big[g(\mathbf{Y},\mathbf{D},\mathbf{X})\,|\,\mathbf{D}_{(1)}>\mathbf{D}_{(0)}\big]=\frac{1}{P(\mathbf{D}_{(1)}>\mathbf{D}_{(0)})}E\big[\kappa_n\cdot g(\mathbf{Y},\mathbf{D},\mathbf{X})\big]$$

成立。其中 $\mathbf{D}_{(1)}=(D_{1(1)},\cdots,D_{n(1)})$, $\mathbf{D}_{(0)}=(D_{1(0)},\cdots,D_{n(0)})$, $\mathbf{D}_{(1)}>\mathbf{D}_{(0)}$ 是指 $D_{1(1)}>D_{1(0)},\cdots,D_{n(1)}>D_{n(0)}$ 同时成立。

特别地，当$n=2$时，即D_1与D_2相关，则有

$$\begin{aligned}&E\big[g(Y_1,Y_2,D_1,D_2,X_1,X_2)\,|\,D_{1(1)}>D_{1(0)},D_{2(1)}>D_{2(0)}\big]\\=\;&\frac{1}{P(D_{1(1)}>D_{1(0)},D_{2(1)}>D_{2(0)})}E\big[\kappa_2\cdot g(Y_1,Y_2,D_1,D_2,X_1,X_2)\big],\end{aligned}$$

其中

$$\begin{aligned}\kappa_2\;=\;&1-\frac{(1-Z_1)(1-Z_2)(1-(1-D_1)(1-D_2))}{P(Z_1=0,Z_2=0\,|\,X_1,X_2)}-\frac{(1-Z_1)Z_2(1-(1-D_1)D_2)}{P(Z_1=0,Z_2=1\,|\,X_1,X_2)}\\&-\frac{Z_1(1-Z_2)(1-D_1(1-D_2))}{P(Z_1=1,Z_2=0\,|\,X_1,X_2)}-\frac{Z_1Z_2(1-D_1D_2)}{P(Z_1=1,Z_2=1\,|\,X_1,X_2)}\end{aligned}$$

成立。

注2.1 此定理对$D_1,\cdots,D_n$相关或不相关都成立，从而推广了Abadi[9]的结果。定理的证明参见附录。

注2.2 当$g(\mathbf{Y},\mathbf{D},\mathbf{X})=1$时，我们能够得到$E[\kappa_n]=P(\mathbf{D}_{(1)}>\mathbf{D}_{(0)})$，所以可以把$\kappa_n$看作权重来识别依从者的期望。定理2.1表明，对依从者来说，条件$\mathbf{X}$下，$\mathbf{D}$具有外源性，不由受试者自身性质决定，所以定理2.1可以用于识别总体中依从组的临床试验参数。

3 模型假设及推断

我们考虑受试者在临床试验过程中被分成m个小组(比如病房)，在第i组里有n_i个受试者。对于第i组里第j个成员，Y_{ij}表示其响应变量即治疗响应结果；X_{ij}表示其p维协变量；$D_{ij(Z_{ij})}$表示其潜在服药状态指示变量，则$D_{ij(1)}$和$D_{ij(0)}$分别表示成员被分到治疗组和对照组后潜在的服药状态；D_{ij}表示其在临床试验中实际服药状态，其中

$$i=1,\cdots,m,\quad j=1,\cdots,n_i,\quad \sum_{i=1}^{m}n_i=n.$$

记

$$\begin{aligned}&\mathbf{D}_i^T=(D_{i1},\cdots,D_{in_i}),\quad \mathbf{D}^T=(\mathbf{D}_1^T,\cdots,\mathbf{D}_m^T),\quad \mathbf{Z}_i^T=(Z_{i1},\cdots,Z_{in_i}),\\&\mathbf{Z}^T=(\mathbf{Z}_1^T,\cdots,\mathbf{Z}_m^T),\quad \mathbf{X}_i^T=(X_{i1},\cdots,X_{in_i}),\quad \mathbf{X}^T=(\mathbf{X}_1^T,\cdots,\mathbf{X}_m^T),\\&\mathbf{Y}_i^T=(Y_{i1},\cdots,Y_{in_i}),\quad \mathbf{Y}^T=(\mathbf{Y}_1^T,\cdots,\mathbf{Y}_m^T),\\&\mathbf{D}_{i(\mathbf{Z}_i)}^T=\big(D_{i1(Z_{i1})},\cdots,D_{in_i(Z_{in_i})}\big),\quad \mathbf{D}_{(\mathbf{Z})}^T=\big(\mathbf{D}_{1(\mathbf{Z}_1)}^T,\cdots,\mathbf{D}_{m(\mathbf{Z}_m)}^T\big).\end{aligned}$$

假设3.1 (i) 同一病房内受试者服药相互影响，不同病房的受试者服药不受影响，即组内D_{ij}与D_{ik}相关，$j\neq k$，组间D_{ij}与D_{mn}独立，$i\neq m$；

(ii) 受试者所服药物对自身治疗响应的作用不受其他受试者影响，即组内$Y_{ij}\,|\,D_{ij}$与$Y_{ik}\,|\,D_{ik}$条件独立，$j\neq k$；

(iii) 设每个成员响应变量与服药量和协变量有如下的模型结构

$$E\big[Y\,|\,D,X,D_{(1)}>D_{(0)}\big]=h(\alpha D+X^T\beta),$$

其中函数 $h(\cdot)$ 二阶连续可导，$h(\cdot)\in[0,1]$；未知参数为 $\theta=(\alpha,\beta^T)^T$，$\alpha$ 反映了药物 D 对治疗响应 Y 的影响，即治疗效果参数，β 反映了受试者自身因素对治疗响应的影响程度。那么 α 即为感兴趣参数，β 为多余参数。

在给定协变量的条件下，记 $P(Y_{ij}=y_{ij}\,|\,D_{ij}=d_{ij},X_{ij}=x_{ij};\alpha,\beta)$ 表示第 i 组第 j 个成员 Y_{ij} 的条件概率，$P(\mathbf{Y}_i=\mathbf{y}_i\,|\,\mathbf{D}_i=\mathbf{d}_i,\mathbf{X}_i=\mathbf{x}_i;\alpha,\beta)$ 表示第 i 组成员 $\mathbf{Y}_i$ 的条件概率，$P(\mathbf{Y}_i=\mathbf{y}_i,\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i=\mathbf{x}_i;\alpha,\beta)$ 表示第 i 组成员 $\mathbf{Y}_i$ 和 $\mathbf{D}_i$ 联合概率，$P(\mathbf{Y}_i=\mathbf{y}_i,\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i=\mathbf{x}_i;\alpha,\beta;\mathbf{D}_{i(1)}>\mathbf{D}_{i(0)})$ 表示依从者中第 i 组成员的联合概率。

利用定理 2.1 得到

$$\begin{aligned}&E\big[\ln P(\mathbf{Y}_i=\mathbf{y}_i,\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i=\mathbf{x}_i;\alpha,\beta)\,|\,\mathbf{D}_{i(1)}>\mathbf{D}_{i(0)}\big]\\ =&\ \frac{1}{P(\mathbf{D}_{i(1)}>\mathbf{D}_{i(0)})}E\big[\kappa_i\ln P(\mathbf{Y}_i=\mathbf{y}_i,\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i=\mathbf{x}_i;\alpha,\beta)\big].\end{aligned}\tag{1}$$

可以用方程的右边代替左边的对数似然函数来构造估计方程。从而，我们可以通过极大化

$$l\big(\alpha,\beta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x}\big)=\frac{1}{m}\sum_{i=1}^{m}\frac{\kappa_{in_i}}{P_{D_i}}\ln P\big(\mathbf{Y}_i=\mathbf{y}_i,\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i=\mathbf{x}_i;\alpha,\beta\big),\tag{2}$$

来估计未知参数。其中 P_{D_i} 是 $P(\mathbf{D}_{i(1)}>\mathbf{D}_{i(0)})$ 的估计

$$P_{D_i}=\frac{\sum_{j=1}^{n_i}d_{ij}z_{ij}}{\sum_{j=1}^{n_i}z_{ij}}-\frac{\sum_{j=1}^{n_i}d_{ij}(1-z_{ij})}{\sum_{j=1}^{n_i}(1-z_{ij})},$$

κ_{in_i} 相应于定理 2.1 中 κ_n，即

$$\kappa_{in_i}=\begin{cases}1, & \mathbf{z}_i=\mathbf{d}_i,\\ 1-\dfrac{1}{\prod\limits_{\substack{z_{ij}=0,1\\ j=1,\cdots,n_i}}\left(\frac{\sum_{j=1}^{n_i}z_{ij}}{n_i}\right)^{z_{ij}}\left(\frac{\sum_{j=1}^{n_i}(1-z_{ij})}{n_i}\right)^{1-z_{ij}}}, & \text{其他},\end{cases}$$

由 (2) 式有

$$\begin{aligned}l\big(\alpha,\beta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x}\big)\ =&\ \sum_{i=1}^{m}\frac{\kappa_{in_i}}{P_{D_i}}\big[\ln P(\mathbf{Y}_i=\mathbf{y}_i\,|\,\mathbf{D}_i=\mathbf{d}_i,\mathbf{X}_i=\mathbf{x}_i;\alpha,\beta)+\ln P(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i=\mathbf{x}_i;\alpha)\big]\\ =&\ \sum_{i=1}^{m}\frac{\kappa_{in_i}}{P_{D_i}}\Big[\sum_{j=1}^{n_i}\big[y_{ij}\ln h(\alpha d_{ij}+x_{ij}^T\beta)+(1-y_{ij})\ln\big(1-h(\alpha d_{ij}+x_{ij}^T\beta)\big)\big]\\ &+\ln P(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i=\mathbf{x}_i;\alpha)\Big].\end{aligned}$$

参数 $\theta=(\alpha,\beta)$ 的估计

$$\begin{aligned}\hat\theta\ =&\ \arg\max_{\theta\in\Theta}\Big\{\sum_{i=1}^{m}\frac{\kappa_{in_i}}{P_{D_i}}\Big[\sum_{j=1}^{n_i}\big[y_{ij}\ln h(\alpha d_{ij}+x_{ij}^T\beta)\\ &+(1-y_{ij})\ln\big(1-h(\alpha d_{ij}+x_{ij}^T\beta)\big)\big]+\ln P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i=\mathbf{x}_i;\alpha\big)\Big]\Big\}.\end{aligned}$$

接下来，利用α的估计，考虑构造α的记分检验。记

$$h_{ij} \triangleq h(\alpha D_{ij} + X_{ij}^T\beta), \quad h'_{ij} \triangleq h'(\alpha D_{ij} + X_{ij}^T\beta), \quad h''_{ij} \triangleq h''(\alpha D_{ij} + X_{ij}^T\beta).$$

定理3.1 在假设3.1下，检验问题$H_0 : \alpha = \alpha_0 \leftrightarrow H_1 : \alpha \neq \alpha_0$的记分检验统计量可表示为$SC(\alpha_0) = U_\alpha(\tilde{\theta})\Sigma^{11}(\tilde{\theta})U_\alpha(\tilde{\theta})$。当

$$n = \sum_{i=1}^{m} n_i \to \infty,$$

且存在$0 < c_0 < c_1 < 1$，使得

$$c_0 \leq \min_i \frac{n_i}{n} \Big/ \max_i \frac{n_i}{n} \leq c_1$$

时，有$SC(\alpha_0) \xrightarrow{L} \chi_1^2$，其中$\theta = (\alpha, \beta^T)^T$，$\tilde{\theta} = (\alpha_0, \tilde{\beta^T})^T$，$\tilde{\beta}$为$H_0$下$\beta$的极大似然估计，即$U_\beta = 0$的解，$\Sigma^{11}$表示$\Sigma^{-1}$第一行第一列元素。$U_\alpha$, U_β, Σ的形式如下

$$U_\alpha = \sum_{i=1}^{m} \frac{\kappa_{in_i}}{P_{D_i}} \Big[\sum_{j=1}^{n_i} \frac{y_{ij}h'_{ij}d_{ij} - h_{ij}h'_{ij}d_{ij}}{h_{ij}(1-h_{ij})} + \frac{\partial \ln P(\mathbf{D}_i = \mathbf{d}_i \,|\, \mathbf{X}_i; \alpha)}{\partial \alpha} \Big],$$

$$U_\beta = \sum_{i=1}^{m} \frac{\kappa_{in_i}}{P_{D_i}} \Big[\sum_{j=1}^{n_i} \frac{y_{ij}h'_{ij}x_{ij} - h_{ij}h'_{ij}x_{ij}}{h_{ij}(1-h_{ij})} \Big],$$

$$\Sigma = \begin{pmatrix} \mathrm{Cov}(U_\alpha, U_\alpha) & \mathrm{Cov}(U_\alpha, U_\beta) \\ \mathrm{Cov}(U_\alpha, U_\beta) & \mathrm{Cov}(U_\beta, U_\beta) \end{pmatrix},$$

其中

$$\mathrm{Cov}(U_\alpha, U_\alpha) = \sum_{i=1}^{m} \sum_{\substack{d_{ij}=0,1 \\ \text{取值全排列}}} \frac{\kappa_i^2}{P_{D_i}^2} \Big[\sum_{j=1}^{n_i} \frac{h'^2_{ij}}{h_{ij}(1-h_{ij})} d_{ij}^2 \Big] P(\mathbf{D}_i = \mathbf{d}_i \,|\, \mathbf{X}_i)$$
$$+ \sum_{i=1}^{m} \sum_{\substack{d_{ij}=0,1 \\ \text{取值全排列}}} \frac{\kappa_i^2}{P_{D_i}^2} \Big[\frac{\partial \ln P(\mathbf{D}_i = \mathbf{d}_i \,|\, \mathbf{X}_i; \alpha)}{\partial \alpha} \Big]^2 P(\mathbf{D}_i = \mathbf{d}_i \,|\, \mathbf{X}_i),$$

$$\mathrm{Cov}(U_\beta, U_\beta) = \sum_{i=1}^{m} \sum_{\substack{d_{ij}=0,1 \\ \text{取值全排列}}} \frac{\kappa_i^2}{P_{D_i}^2} \Big[\sum_{j=1}^{n_i} \frac{h'^2_{ij}}{h_{ij}(1-h_{ij})} x_{ij}x_{ij}^T \Big] P(\mathbf{D}_i = \mathbf{d}_i \,|\, \mathbf{X}_i),$$

$$\mathrm{Cov}(U_\alpha, U_\beta) = \sum_{i=1}^{m} \sum_{\substack{d_{ij}=0,1 \\ \text{取值全排列}}} \frac{\kappa_i^2}{P_{D_i}^2} \Big[\sum_{j=1}^{n_i} \frac{h'^2_{ij}}{h_{ij}(1-h_{ij})} d_{ij}x_{ij} \Big] P(\mathbf{D}_i = \mathbf{d}_i \,|\, \mathbf{X}_i).$$

给定检验水平γ，近似拒绝域为$R^+ = \{SC(\alpha_0) > \chi_1^2(1-\gamma)\}$。

注3.1 定理中记分检验统计量渐近方差的计算见附录，渐近分布的证明与一般记分检验统计量渐近分布证明类似，略。

4 Logit 模型下的统计推断

Logit 模型是一个常见的模型，也称为 Logistic 回归模型，本节针对 Logit 模型和受试者服药状态相关性的假定下，计算检验统计量。

假设 4.1 在假设 3.1 成立的条件下，进一步有

(i) 设每个成员响应变量与服药量和协变量有如下的模型结构

$$\text{Logit}\ \ E\big[Y\,|\,D,X,D_{(1)}>D_{(0)}\big]=\alpha D+X^T\beta,$$

其等价形式为

$$E\big[Y\,|\,D,X,D_{(1)}>D_{(0)}\big]=\frac{\mathrm{e}^{\alpha D+X^T\beta}}{1+\mathrm{e}^{\alpha D+X^T\beta}},$$

其中 α 反映了药物 D 对治疗响应 Y 的影响，即治疗效果参数，β 反映了受试者自身因素对治疗响应的影响程度。

(ii) 设第 i 组各分量取 1 的概率相等且与协变量无关，即

$$P(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i)=P(\mathbf{D}_i=\mathbf{d}_i),\quad P(D_{i1}=1)=P(D_{i2}=1)=\cdots=P(D_{in}=1)=q,$$

且 q 有先验分布

$$q\sim\frac{p^{\mathrm{e}^{\alpha}+a-1}+(1-p)^{\mathrm{e}^{\alpha}+b-1}}{\mathrm{Be}(\mathrm{e}^{\alpha}+a,\mathrm{e}^{\alpha}+b)},\qquad p\in(0,1),$$

上述假设表明，不仅同一组内个体依从性是相关的，而且个体依从性可以依赖于治疗效果。

由假设 4.1 的 (ii) 可推出

$$\begin{aligned}P_{i0}&\triangleq P(\mathbf{D}_i=\mathbf{d}_i)\\&=\int_0^1 C_{n_i}^{\sum_{j=1}^{n_i}d_{ij}}p^{\sum_{j=1}^{n_i}d_{ij}}(1-p)^{n_i-\sum_{j=1}^{n_i}d_{ij}}\frac{p^{\mathrm{e}^{\alpha}+a-1}(1-p)^{\mathrm{e}^{\alpha}+b-1}}{\mathrm{Be}(\mathrm{e}^{\alpha}+a,\mathrm{e}^{\alpha}+b)}\mathrm{d}p\\&=\frac{C_{n_i}^{\sum_{j=1}^{n_i}d_{ij}}\mathrm{Be}\big(\mathrm{e}^{\alpha}+a+\sum_{j=1}^{n_i}d_{ij},\mathrm{e}^{\alpha}+b+n_i-\sum_{j=1}^{n_i}d_{ij}\big)}{\mathrm{Be}(\mathrm{e}^{\alpha}+a,\mathrm{e}^{\alpha}+b)}\\&=\frac{C_{n_i}^{\sum_{j=1}^{n_i}d_{ij}}\prod_{r_1=0}^{\sum_{j=1}^{n_i}d_{ij}-1}(\mathrm{e}^{\alpha}+a+r_1)\prod_{r_2=0}^{n_i-\sum_{j=1}^{n_i}d_{ij}-1}(\mathrm{e}^{\alpha}+b+r_2)}{\prod_{r_3=0}^{n_i-1}(2\mathrm{e}^{\alpha}+a+b+r_3)}.\end{aligned}$$

计算 $\ln P(\mathbf{D}_i=\mathbf{d}_i)$ 对各变量的一阶导数

$$\begin{aligned}P_{i\alpha}&\triangleq\frac{\partial\ln P(\mathbf{D}_i=\mathbf{d}_i)}{\partial\alpha}\\&=\sum_{r_1=0}^{\sum_{j=1}^{n_i}d_{ij}-1}\Big(\frac{\mathrm{e}^{\alpha}}{\mathrm{e}^{\alpha}+a+r_1}\Big)+\sum_{r_2=0}^{n_i-\sum_{j=1}^{n_i}d_{ij}-1}\Big(\frac{\mathrm{e}^{\alpha}}{\mathrm{e}^{\alpha}+b+r_2}\Big)-\sum_{r_3=0}^{n_i-1}\Big(\frac{2\mathrm{e}^{\alpha}}{2\mathrm{e}^{\alpha}+a+b+r_3}\Big),\end{aligned}$$

$$P_{ia}\triangleq\frac{\partial\ln P(\mathbf{D}_i=\mathbf{d}_i)}{\partial a}=\sum_{r_1=0}^{\sum_{j=1}^{n_i}d_{ij}-1}\Big(\frac{1}{\mathrm{e}^{\alpha}+a+r_1}\Big)-\sum_{r_3=0}^{n_i-1}\Big(\frac{1}{2\mathrm{e}^{\alpha}+a+b+r_3}\Big),$$

$$P_{ib}\triangleq\frac{\partial\ln P(\mathbf{D}_i=\mathbf{d}_i)}{\partial b}=\sum_{r_2=0}^{n_i-\sum_{j=1}^{n_i}d_{ij}-1}\Big(\frac{1}{\mathrm{e}^{\alpha}+b+r_2}\Big)-\sum_{r_3=0}^{n_i-1}\Big(\frac{1}{2\mathrm{e}^{\alpha}+a+b+r_3}\Big).$$

定理 4.1 在假设 4.1 下，假设检验 $H_0:\alpha=\alpha_0\leftrightarrow H_1:\alpha\neq\alpha_0$ 的记分检验统计量可表示为 $SC_1(\alpha_0)=U_\alpha(\tilde{\theta})\Sigma^{11}(\tilde{\theta})U_\alpha(\tilde{\theta})$。当

$$n=\sum_{i=1}^{m}n_i\to\infty,$$

且存在 $0<c_0<c_1<1$，使得

$$c_0\le\min_i\frac{n_i}{n}\Big/\max_i\frac{n_i}{n}\le c_1$$

时，有 $SC_1(\alpha_0)\xrightarrow{L}\chi_1^2$，其中 $\theta=(\alpha,\beta^T,a,b)^T$，$\tilde{\theta}=(\alpha_0,\tilde{\beta^T},\tilde{a},\tilde{b})^T$，$\tilde{\beta}$, $\tilde{a}$, $\tilde{b}$ 分别为 H_0 下 β, a, b 的极大似然估计，即分别为 $U_\beta=0$, $U_a=0$, $U_b=0$ 的解，Σ^{11} 表示 Σ^{-1} 第一行第一列元素。U_α, U_β, U_a, U_b, Σ 形式如下

$$U_\alpha=\sum_{i=1}^{m}\frac{\kappa_{in_i}}{P_{D_i}}\Big[\sum_{j=1}^{n_i}\Big(y_{ij}d_{ij}-\frac{\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta}}{1+\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta}}d_{ij}\Big)+P_{i\alpha}\Big],$$

$$U_\beta=\sum_{i=1}^{m}\frac{\kappa_{in_i}}{P_{D_i}}\Big[\sum_{j=1}^{n_i}\Big(y_{ij}x_{ij}-\frac{\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta}}{1+\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta}}x_{ij}\Big)\Big],$$

$$U_a=\sum_{i=1}^{m}\frac{\kappa_{in_i}}{P_{D_i}}P_{ia},\quad U_b=\sum_{i=1}^{m}\frac{\kappa_{in_i}}{P_{D_i}}P_{ib},$$

$$\Sigma=\begin{pmatrix}\mathrm{Cov}(U_\alpha,U_\alpha)&\mathrm{Cov}(U_\alpha,U_\beta)&\mathrm{Cov}(U_\alpha,U_a)&\mathrm{Cov}(U_\alpha,U_b)\\\mathrm{Cov}(U_\alpha,U_\beta)&\mathrm{Cov}(U_\beta,U_\beta)&0&0\\\mathrm{Cov}(U_\alpha,U_a)&0&\mathrm{Cov}(U_a,U_a)&\mathrm{Cov}(U_a,U_b)\\\mathrm{Cov}(U_\alpha,U_b)&0&\mathrm{Cov}(U_a,U_b)&\mathrm{Cov}(U_b,U_b)\end{pmatrix},$$

其中

$$\mathrm{Cov}(U_\alpha,U_\alpha)=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_{in_i}^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta}}{(1+\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta})^2}d_{ij}^2+P_{i\alpha}^2\Big]P_{i0},$$

$$\mathrm{Cov}(U_\alpha,U_\beta)=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_{in_i}^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta}}{(1+\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta})^2}d_{ij}x_{ij}\Big]P_{i0},$$

$$\mathrm{Cov}(U_\alpha,U_a)=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_{in_i}^2}{P_{D_i}^2}P_{i\alpha}P_{ia}P_{i0},$$

$$\mathrm{Cov}(U_\alpha,U_b)=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_{in_i}^2}{P_{D_i}^2}P_{i\alpha}P_{ib}P_{i0},$$

$$\mathrm{Cov}(U_\beta,U_\beta)=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_{in_i}^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta}}{(1+\mathrm{e}^{\alpha d_{ij}+x_{ij}^T\beta})^2}x_{ij}x_{ij}^T\Big]P_{i0},$$

$$\mathrm{Cov}(U_a,U_a)=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_{in_i}^2}{P_{D_i}^2}P_{ia}^2P_{i0},$$

$$\mathrm{Cov}(U_a,U_b)=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_{in_i}^2}{P_{D_i}^2}P_{ia}P_{ib}P_{i0},$$

$$\mathrm{Cov}(U_b,U_b)=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_{in_i}^2}{P_{D_i}^2}P_{ib}^2P_{i0}.$$

给定检验水平 γ，近似拒绝域为 $R^+=\{SC_1(\alpha_0)>\chi_1^2(1-\gamma)\}$。

5 模拟研究

针对上述结论进行模拟研究，考察检验方法的优劣。

1) 设定真值产生随机数 $(X_{ij},Z_{ij},D_{ij},Y_{ij})$，$i=1,\cdots,m$，$j=1,\cdots,n$。

选定参数真值 $\alpha_0=0$，$\beta=0.5$，$a=4$，$b=4$，产生随机数 $Z_{ij}\sim\mathrm{B}(1,0.5)$，$X_{ij}\sim\mathrm{U}(-0.5,0.5)$，$P_{D_i}\sim\mathrm{Be}(\mathrm{e}^\alpha+a,\mathrm{e}^\alpha+b)$，由 P_{D_i} 产生随机数 $D_{ij}\sim\mathrm{B}(1,P_{D_i})$，代入模型

$$P_{Y_{ij}}=P\big(Y_{ij}=1\,|\,D_{ij},X_{ij},D_{ij(1)}>D_{ij(0)}\big)=\frac{\mathrm{e}^{\alpha D_{ij}+\beta X_{ij}}}{1+\mathrm{e}^{\alpha D_{ij}+\beta X_{ij}}},$$

产生随机数 $Y_{ij}\sim\mathrm{B}(1,P_{Y_{ij}})$。

2) 计算检验统计量 SC_1 和拒绝 H_0 的概率。

求解方程组 $U_\beta(0,\beta,a,b)=0$，$U_a(0,\beta,a,b)=0$，$U_b(0,\beta,a,b)=0$，解得 $\tilde\beta,\tilde a,\tilde b$ 分别为 H_0 下 β,a,b 的极大似然估计，计算记分检验统计量，进而得到拒绝 H_0 的概率。通过5000次模拟，结果如表1所示，包括 $\alpha=0$ 下 β,a,b 的参数估计以及 $\alpha=0$ 下拒绝 H_0 的概率。

表1: $\alpha=0$ 下参数估计和假设检验模拟结果

样本量		参数估计			假设检验	
组内 n	组间 m	$\hat\beta(\mathrm{MSE}(\hat\beta))$	$\hat a(\mathrm{MSE}(\hat a))$	$\hat b(\mathrm{MSE}(\hat b))$	$\gamma=0.05$	$\gamma=0.01$
5	5	0.5757(1.7902)	4.1939(0.5971)	4.1963(0.5957)	0.0454	0.0109
	10	0.5460(1.2518)	4.0904(0.1693)	4.0887(0.1706)	0.0455	0.0070
	20	0.5161(0.6559)	4.0210(0.0819)	4.0225(0.0832)	0.0480	0.0096
	50	0.5121(0.2510)	4.0008(0.0334)	4.0011(0.0339)	0.0482	0.0100
10	5	0.5648(1.2633)	4.2275(0.3076)	4.2262(0.3050)	0.0469	0.0108
	10	0.5143(0.5706)	4.0982(0.1657)	4.0962(0.1639)	0.0490	0.0096
	20	0.5097(0.3104)	4.0476(0.0814)	4.0497(0.0812)	0.0510	0.0122
	50	0.5090(0.1318)	4.0182(0.0326)	4.0187(0.0327)	0.0494	0.0096

设定组内成员数 $n = 5$，组数 $m = 20$，给定检验水平 $\gamma = 0.05$ 和 $\gamma = 0.01$，通过1000次模拟，计算功效结果如图1和表2所示，图1直观给出功效值随 α 取值的变化趋势，表2进一步给出 α 取值对应的功效函数值。

表1表明，针对服药状态相关情况下，记分检验方法能够控制犯第一类错误的概率；尤其在总样本量不大的情况下，也能较好的控制水平；随着组数和组内成员数的增大，犯第一类错误的概率接近给定的检验水平。另外，模拟结果显示，随着组数和组内成员数的增加，参数估计的偏差及均方误差 (MSE) 迅速减小。从图1和表2看出，该检验方法有很好的功效，当 α 接近 ±1.5 时，功效接近1。

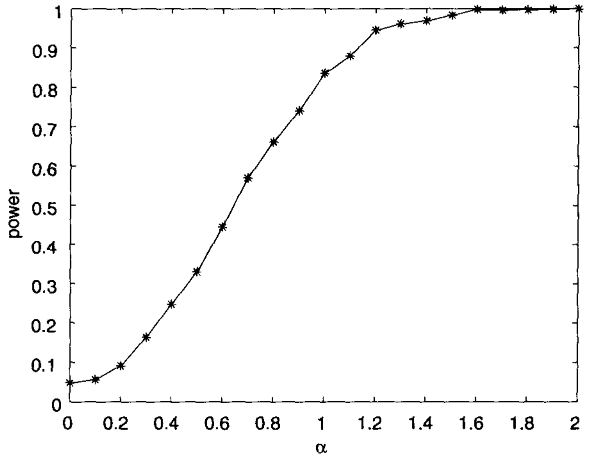

图 1: 假设检验方法功效函数图 (检验水平 0.05)

表 2: 功效函数值表

检验水平 $\gamma = 0.05$				检验水平 $\gamma = 0.01$			
α	功效值	α	功效值	α	功效值	α	功效值
-2	0.998	0.1	0.056	-2	0.990	0.1	0.016
-1.5	0.982	0.2	0.091	-1.5	0.934	0.2	0.021
-1.2	0.930	0.3	0.163	-1.2	0.800	0.3	0.049
-1	0.839	0.4	0.247	-1	0.647	0.4	0.090
-0.9	0.770	0.5	0.330	-0.9	0.534	0.5	0.145
-0.8	0.671	0.6	0.444	-0.8	0.413	0.6	0.223
-0.7	0.574	0.7	0.570	-0.7	0.339	0.7	0.341
-0.6	0.450	0.8	0.661	-0.6	0.217	0.8	0.399
-0.5	0.322	0.9	0.740	-0.5	0.135	0.9	0.501
-0.4	0.259	1	0.835	-0.4	0.116	1	0.618
-0.3	0.171	1.2	0.945	-0.3	0.048	1.2	0.812
-0.2	0.109	1.5	0.983	-0.2	0.032	1.5	0.927
-0.1	0.067	2	0.999	-0.1	0.023	2	0.996

6 结论与讨论

本文考虑受试者服药状态相关的情况，并且服药状态变量D的分布与治疗参数有关，同时考虑协变量对治疗效果有影响这一因素。首先给出在该情况下治疗效果参数可识别的结论，接下来利用极大似然估计构造记分检验统计量，该检验方法在满足假设2.1的条件下，进一步假设给定D, X下Y的条件概率形式以及D的分布形式。模拟研究中，给定特殊模型进行模拟，结果表明，该假设检验方法在大样本的情况下，犯第一类错误的概率接近给定的检验水平；样本量不大的情况下，也能够控制犯第一类错误的概率，并接近给定的检验水平；当组内成员数和组数增大时，检验效果更为理想。

7 附录

定理2.1的证明 由于我们考虑的是违背者不存在的情况，即$D_{i(1)} \geq D_{i(0)}$，$i=1,\cdots,n$，因此D_i的取值为0, Z_i, 1。那么$D_1,\cdots,D_n$取值全排列共3^n项，其中一排列$D_i = Z_i$，$i = 1,\cdots,n$等价于$\mathbf{D}_{(1)} > \mathbf{D}_{(0)}$成立。

记$g(\cdot) = g(Y_1,\cdots,Y_n,D_1,\cdots,D_n,X_1,\cdots,X_n)$，则有下式成立

$$\begin{aligned}
& E\big[g(\cdot)\,|\,\mathbf{X},\mathbf{D}_{(1)} > \mathbf{D}_{(0)}\big] \\
= & \frac{1}{P(\mathbf{D}_{(1)} > \mathbf{D}_{(0)}\,|\,\mathbf{X})}\Big\{E\big[g(\cdot)\,|\,\mathbf{X}\big] \\
& - \sum_{\substack{z_i=0,1\\ i=1,\cdots,n}} \sum_{\substack{d_i=0,z_i,1\text{ 取值全排列}\\ \forall\text{ 排列},\,\exists\, i,d_i\neq z_i}} E\big[g(\cdot)\,|\,\mathbf{X},\mathbf{Z}=\mathbf{z},D_1=d_1,\cdots,D_n=d_n\big] \\
& \times P\big(D_1=d_1,\cdots,D_n=d_n\,|\,\mathbf{X},\mathbf{Z}=\mathbf{z}\big)\Big\}.
\end{aligned} \tag{3}$$

根据假定条件，在给定$\mathbf{X}$条件下，$\mathbf{Z}$与$(\mathbf{Y},\mathbf{D})$独立，当$\mathbf{Z}=\mathbf{z}=(z_1,\cdots,z_n)^T$时

$$\begin{aligned}
& \sum_{\substack{d_i=0,z_i,1\text{取值全排列}\\ \forall\text{ 排列},\exists\, i,d_i\neq z_i}} E\big[g(\cdot)\,|\,\mathbf{X},D_1=d_1,\cdots,D_n=d_n,Z_1=z_1,\cdots,Z_n=z_n\big] \\
& \times P\big(D_1=d_1,\cdots,D_n=d_n\,|\,\mathbf{X},Z_1=z_1,\cdots,Z_n=z_n\big) \\
= & \frac{1}{P(Z_1=z_1,\cdots,Z_n=z_n\,|\,\mathbf{X})} \\
& \times E\Big[\Big(1-\prod_{i=1}^n D_i^{z_i}(1-D_i)^{1-z_i}\Big)\prod_{i=1}^n Z_i^{z_i}(1-Z_i)^{1-z_i}g(\cdot)\,|\,\mathbf{X}\Big].
\end{aligned} \tag{4}$$

注 对(4)式成立进行解释说明。固定Z_i的取值，D_i, $i=1,\cdots,n$的取值全排列中不包括$D_i=Z_i$，$i=1,\cdots,n$这一排列，因此

$$\sum_{\substack{d_i=0,z_i,1\text{取值全排列}\\ \forall\text{ 排列},\exists\, i,d_i\neq z_i}} P\big(D_1=d_1,\cdots,D_n=d_n\,|\,\mathbf{X},Z_1=z_1,\cdots,Z_n=z_n\big) = 1-\prod_{i=1}^n d_i^{z_i}(1-d_i)^{1-z_i}.$$

将 (4) 式代入 (3) 式，两边对 $\mathbf{X}$ 积分得到

$$\int E[g(\cdot)\,|\,\mathbf{X},\mathbf{D}_{(1)}>\mathbf{D}_{(0)}]\mathrm{d}P(\mathbf{X}\,|\,\mathbf{D}_{(1)}>\mathbf{D}_{(0)})$$

$$=\frac{1}{P(\mathbf{D}_{(1)}>\mathbf{D}_{(0)})}\int E\Big[g(\cdot)\Big(1-\sum_{\substack{z_i=0,1\\ i=1,\cdots,n}}\frac{(1-\prod_{i=1}^n D_i^{z_i}(1-D_i)^{1-z_i})\prod_{i=1}^n Z_i^{z_i}(1-Z_i)^{1-z_i}}{P(Z_1=z_1,\cdots,Z_n=z_n\,|\,\mathbf{X})}\Big)\Big|\mathbf{X}\Big]\mathrm{d}P(\mathbf{X}),$$

即可推出

$$\big[g(\mathbf{Y},\mathbf{D},\mathbf{X})\,|\,\mathbf{D}_{(1)}>\mathbf{D}_{(0)}\big]=\frac{1}{P(\mathbf{D}_{(1)}>\mathbf{D}_{(0)})}E\big[\kappa_n\cdot g(\mathbf{Y},\mathbf{D},\mathbf{X})\big],$$

其中

$$\kappa_n=1-\sum_{\substack{z_i=0,1\\ i=1,\cdots,n}}\frac{\prod_{i=1}^n Z_i^{z_i}(1-Z_i)^{1-z_i}(1-\prod_{i=1}^n D_i^{z_i}(1-D_i)^{1-z_i})}{P(Z_1=z_1,\cdots,Z_n=z_n\,|\,X_1,\cdots,X_n)}.$$

定理 3.1 中协方差的计算　首先计算得分函数

$$U(\theta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x})=\begin{pmatrix}U_\alpha(\alpha,\beta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x})\\ U_\beta(\alpha,\beta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x})\end{pmatrix}$$

$$=\sum_{i=1}^m\frac{\kappa_{in_i}}{P_{D_i}}\sum_{j=1}^{n_i}\Big[\frac{(y_{ij}-h_{ij})h'_{ij}}{h_{ij}(1-h_{ij})}\begin{pmatrix}d_{ij}\\ x_{ij}\end{pmatrix}\Big]+\sum_{i=1}^m\frac{\kappa_{in_i}}{P_{D_i}}\begin{pmatrix}\partial\ln P(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha)/\partial\alpha\\ 0\end{pmatrix}.$$

接下来计算协方差阵。$U(\theta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x})$ 可以写为独立和的形式，即

$$U\big(\theta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x}\big)=\sum_{i=1}^n U_i\big(\theta;\mathbf{y}_i,\mathbf{d}_i\,|\,\mathbf{x}_i\big),$$

其中

$$U_i(\theta;\mathbf{y}_i,\mathbf{d}_i\,|\,\mathbf{x}_i)=\begin{pmatrix}U_{i,\alpha}(\alpha,\beta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x})\\ U_{i,\beta}(\alpha,\beta;\mathbf{y},\mathbf{d}\,|\,\mathbf{x})\end{pmatrix}$$

$$=\frac{\kappa_{in_i}}{P_{D_i}}\sum_{j=1}^{n_i}\Big[\frac{(y_{ij}-h_{ij})h'_{ij}}{h_{ij}(1-h_{ij})}\begin{pmatrix}d_{ij}\\ x_{ij}\end{pmatrix}\Big]+\frac{\kappa_{in_i}}{P_{D_i}}\begin{pmatrix}\partial\ln P(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha)/\partial\alpha\\ 0\end{pmatrix},$$

且 U_i，$i=1,\cdots,m$ 相互独立，得到 $\mathrm{Cov}[U(\theta),U(\theta)]=\sum_{i=1}^m\mathrm{Cov}[U_i(\theta),U_i(\theta)]$。在 $H_0:\alpha=\alpha_0$ 下，令 $\beta=\tilde\beta$，则有 $E[U_i(\theta;\mathbf{Y}_i,\mathbf{D}_i\,|\,\mathbf{X}_i=\mathbf{x}_i)]=0$ 成立。得到

$$\mathrm{Cov}\big[U_{i,\alpha}(\theta;\mathbf{Y}_i,\mathbf{D}_i\,|\,\mathbf{X}_i=\mathbf{x}_i),U_{i,\alpha}(\theta;\mathbf{Y}_i,\mathbf{D}_i\,|\,\mathbf{X}_i=\mathbf{x}_i)\big]$$

$$=E_{\mathbf{Y}_i,\mathbf{D}_i}\Big\{\frac{\kappa_i^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{(y_{ij}-h_{ij})h'_{ij}D_{ij}}{h_{ij}(1-h_{ij})}\Big]^2+\frac{\kappa_i^2}{P_{D_i}^2}\Big[\frac{\partial\ln P(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha)}{\partial\alpha}\Big]^2$$

$$+2\frac{\kappa_i^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{(y_{ij}-h_{ij})h'_{ij}D_{ij}}{h_{ij}(1-h_{ij})}\Big]\frac{\partial\ln P(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha)}{\partial\alpha}\Big\}$$

$$\triangleq I_1+I_2+I_3.$$

接下来分别计算 $I_1 - I_3$。

$$I_1 = E_{\mathbf{Y}_i,\mathbf{D}_i}\Big\{\frac{\kappa_i^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{(Y_{ij}-h_{ij})h'_{ij}D_{ij}}{h_{ij}(1-h_{ij})}\Big]^2\Big\} = E_{\mathbf{D}_i}\Big[\frac{\kappa_i^2}{P_{D_i}^2}\sum_{j=1}^{n_i}\frac{h'^2_{ij}}{h_{ij}(1-h_{ij})}D_{ij}^2\,\Big|\,\mathbf{D}_i\Big]$$

$$= \sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_i^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{h'^2_{ij}}{h_{ij}(1-h_{ij})}d_{ij}^2\Big]P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i\big),$$

$$I_2 = E_{\mathbf{Y}_i,\mathbf{D}_i}\Big\{\frac{\kappa_i^2}{P_{D_i}^2}\Big[\frac{\partial\ln P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha\big)}{\partial\alpha}\Big]^2\Big\}$$

$$= \sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_i^2}{P_{D_i}^2}\Big[\frac{\partial\ln P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha\big)}{\partial\alpha}\Big]^2P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i\big),$$

$$I_3 = E_{\mathbf{Y}_i,\mathbf{D}_i}\Big\{2\frac{\kappa_i^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{(y_{ij}-h_{ij})h'_{ij}D_{ij}}{h_{ij}(1-h_{ij})}\Big]\frac{\partial\ln P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha\big)}{\partial\alpha}\Big\}$$

$$= E_{\mathbf{D}_i}\Big\{2\frac{\kappa_i^2}{P_{D_i}^2}\cdot E_{\mathbf{Y}_i}\Big[\Big(\sum_{j=1}^{n_i}\frac{(y_{ij}-h_{ij})h'_{ij}D_{ij}}{h_{ij}(1-h_{ij})}\Big)\,\Big|\,\mathbf{D}_i\Big]\cdot\frac{\partial\ln P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha\big)}{\partial\alpha}\Big\} = 0.$$

因此

$$\begin{aligned}&\mathrm{Cov}\big[U_\alpha(\theta;\mathbf{Y},\mathbf{D}\,|\,\mathbf{X}),U_\alpha(\theta;\mathbf{Y},\mathbf{D}\,|\,\mathbf{X})\big]\\ &=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_i^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{h'^2_{ij}}{h_{ij}(1-h_{ij})}d_{ij}^2\Big]P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i\big)\\ &\quad+\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_i^2}{P_{D_i}^2}\Big[\frac{\partial\ln P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i;\alpha\big)}{\partial\alpha}\Big]^2P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i\big),\end{aligned}$$

同理推出

$$\begin{aligned}&\mathrm{Cov}\big[U_\beta(\theta;\mathbf{Y},\mathbf{D}\,|\,\mathbf{X}),U_\beta(\theta;\mathbf{Y},\mathbf{D}\,|\,\mathbf{X})\big]\\ &=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_i^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{h'^2_{ij}}{h_{ij}(1-h_{ij})}x_{ij}x_{ij}^T\Big]P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i\big),\end{aligned}$$

$$\begin{aligned}&\mathrm{Cov}\big[U_\alpha(\theta;\mathbf{Y},\mathbf{D}\,|\,\mathbf{X}),U_\beta(\theta;\mathbf{Y},\mathbf{D}\,|\mathbf{X})\big]\\ &=\sum_{i=1}^{m}\sum_{\substack{d_{ij}=0,1\\ \text{取值全排列}}}\frac{\kappa_i^2}{P_{D_i}^2}\Big[\sum_{j=1}^{n_i}\frac{h'^2_{ij}}{h_{ij}(1-h_{ij})}d_{ij}x_{ij}\Big]P\big(\mathbf{D}_i=\mathbf{d}_i\,|\,\mathbf{X}_i\big).\end{aligned}$$

参考文献：

[1] Imbens G W, Angrist J D. Indentification and estimation of local average treatment effect[J]. Econometrica, 1994, 62: 467-476

[2] Angrist J D, *et al.* Identification of causal effects using instrumental variables[J]. Journal of the American Statistical Association, 1996, 91: 444-472

[3] Angrist J D, Imbens G W. Two-stage least squares estimation of average causal effects in models with variable treatment intensity[J]. Journal of the American Statistical Association, 1995, 90: 431-442

[4] Robins J M, Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models[J]. Biometrika, 2004, 56: 763-783

[5] Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models[J]. J R Statist Soc B, 2003, 65: 817-835

[6] Imbens G W, Rubin D B. Estimating outcome distributions for compliers in instrumental variable models[J]. Review of Economic Studies, 1997, 64: 555-574

[7] Abadie A. Bootstrap test for distributional treatment effects in instrumental variable models[J]. Journal of the American Statistical Association, 2002, 97: 284-292

[8] Abadie A, Angrist J D, Imbens G W. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings[J]. Econometrica, 2002, 70: 91-117

[9] Abadie A. Semiparametric instrumental variable estimation of treatment response models[J]. Journal Econometric, 2003, 113: 231-263

Inference of Treatment Effects Based on Correlated Data from Randomized Trials with Non-compliance

DAI Jun,　ZHANG Zhong-zhan

(College of Applied Sciences, Beijing University of Technology, Beijing 100124)

Abstract: Assume that in a trial, the compliances among patients are correlated, and the co-variables of patients may affect the evaluation of the treatment. In this paper, we prove the identifiability of the treatment effect. Based on an MLE of parameters, a score test is constructed for the treatment effect. Simulation results show that the score test can control type I error even for small samples.

Keywords: non-compliance; instrumental variable; correlated data; score test

Received: 14 July 2008.　**Accepted:** 09 Apr 2009.

Foundation item: The Natural Science Foundation of Beijing City (1072003); the National Natural Science Foundation of China (10971007).